Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5029–5034
Marseille, 20-25 June 2022

Huqariq: A Multilingual Speech Corpus of Native Languages of Peru for Speech Recognition

Rodolfo Zevallos, Luis Camacho, Nelsi Melgarejo
Pompeu Fabra University, Pontifical Catholic University of Peru
Barcelona Spain, Lima Peru
rodolfojoel.zevallos@upf.edu, {luis.camacho, nelsi.melgarejo}@pucp.pe

Abstract

The Huqariq corpus is a multilingual collection of speech from native Peruvian languages. The transcribed corpus is intended for the research and development of speech technologies to preserve endangered languages in Peru. Huqariq is primarily designed for the development of automatic speech recognition, language identification and text-to-speech tools. In order to achieve corpus collection sustainably, we employ the crowdsourcing methodology. Huqariq includes four native languages of Peru, and it is expected that by the end of the year 2022, it can reach up to 20 native languages out of the 48 native languages in Peru. The corpus has 220 hours of transcribed audio recorded by more than 500 volunteers, making it the largest speech corpus for native languages in Peru. In order to verify the quality of the corpus, we present speech recognition experiments using 220 hours of fully transcribed audio.

Keywords: Speech Corpus, Speech Recognition, Low-resource Languages

1. Introduction

The Huqariq project responds to the endangerment currently faced by native languages in Latin America and the lack of language technologies faced by low-resource languages in Peru (Rogers and Campbell, 2015). This situation is mainly due to the lack of speech corpora, which are the raw material for the creation of language tools, are scarce and the few that exist are privately licensed; for this reason, they should be in the public domain to contribute to the development and revitalization of languages.

Around the world, there are some initiatives for the collection of corpora for low-resources languages and that are in the public domain that employ different methodologies and ways of collection. One of the most successful methodologies is crowdsourcing, i.e. native speakers volunteering to help in the construction of the resources. This methodology is supported by web tools or mobile applications, which can be massively used.

Our corpus collection tool is designed to expand organically to new native languages as community members record domain-specific base audios as prompts in the corpus collection. Unlike others (Common Voice (Ardila et al., 2019)), this tool does not use text to be read, as many native speakers of indigenous languages are illiterate in their native language. Therefore, our tool replaces reading texts with listening to audios. This subtle but important change facilitates corpus collection.

2. Prior work

Although the majority of the speech corpora employed in the most widely used tools are private, there are some worldwide initiatives of speech corpora with open licenses and low-resource languages. In 2019 the VoxForge project (VoxForge, 2019) collected a speech corpus for 17 languages; this project is community-driven in the same way Mozilla's Common Voice project (Ardila et al., 2019) has collected 2500 hours of transcribed audio for 29 languages by crowdsourcing being one of the most community-supported projects.

On the other hand, the speech corpus projects of native languages of Latin America are almost null; in 2019, a speech corpus of 142 hours of fully transcribed Mapudungun was released (Duan et al., 2019); in 2018, the siminchikkunarayku (Cardenas et al., 2018) project collected 99 hours of audio of southern Quechua. Unfortunately, both projects do not have open licenses.

3. Native Languages

Peru is a multicultural country, mainly due to the presence of native first nations, these make up a total of 10% of the population. Because of this population there are still 48 native languages spoken, however, they are under the risk of extinction. These languages are facing some major issues like the lack of a unique grammar or writing system, lack of presence on the internet, lack of mass of expert linguists and lack of electronic resources. (Cardenas et al., 2018) In this section we present some important linguistic characteristics relevant to NLP, especially in regards to dialectal and phonological variety which play an important role in speech based linguistic technology.

3.1. Quechua

Quechua (ISO 639-3 que) is a family of languages spoken in South America with about 10 million speakers, not only in the Andean regions but also in the valleys and plains connecting the Amazon jungle and the Pacific coast. Quechua languages are considered highly agglutinative with a subject-object-verb (SOV) sentence structure as well as mostly postpositional. Even though the classification of Quechua languages remains open to research (Heggarty et al., 2005; Landerman, 1992), recent work in language technology for Quechua (Rios, 2015; Rios and Mamani, 2014) have adopted the categorization system described by Torero (Torero, 1964). This categorization divides the Quechua languages into two main branches, QI (Glottolog quec1386) and QII (quec1388). Branch QI corresponds to the dialects spoken in central Peru, which are treated as one collective in this paper. QII is further divided in three branches, QIIA, QIIB and QIIC. QIIA groups the dialects spoken in Northern Peru, while QIIB the ones in Ecuador and Colombia. In this paper we work with QI (Central Quechua, Glottolog quec1386) and QIIC (Southern Quechua, Glottolog quec1389).

3.1.1. Southern Quechua

Southern Quechua (QIIC) has two main variants: Chanka Quechua (ISO 639-3 quy) and Collao Quechua, also known as Cusco Quechua (ISO 639-3 quz). In both dialects, only the vowels /a/, /i/ and /u/ are found as phonemic vowels. Referring to consonants, Chanka Quechua has a total of 15, most of them voiceless and as in Spanish, the phoneme /t͜ʃ/ is written as *ch*, /ɲ/ as *ñ*, and /ʎ/ as *ll*. On the other hand, Collao Quechua also has a glottal and an aspirated version of each plosive consonant, giving it a total of 25 consonants. Both dialects have voiced consonants in their phonemic inventory due to the large number of borrowings from Spanish.

3.1.2. Central Quechua

Since there is greater dialectal variation among the variants of the QI branch compared to the variation between Quechua Chanka and Quechua Collao, we will go into a bit more detail in this section. Unlike Southern Quechua, Central Quechua (QI) has 3 short phonemic vowels /a/, /i/ and /u/, and 3 long phonemic vowels /aa/, /ii/ and /uu/. Central Quechua also has 3 nasal consonants /m/ /n/, /ɲ/, 4 occlusive consonants /p/, /t/, /k/, /q/, 2 affricate consonants of variable value, 3 fricative consonants /s/, /ʃ/, /h/, 2 approximant consonants /j/, /w/ and 3 liquid consonants /λ/, /ɾ/, /l/. The uvular /q/ is pronounced occlusive only in Callejón de Huaylas, being fricative in the other provinces: voiceless [χ] in Corongo for all positions, while in the Conchucos it is voiced [ʁ] in initial position and voiceless in coda. The alveolar nasal /n/ has three allophones, namely: the velar [ɲ], in syllabic coda and when it precedes the velar [k], the uvular [ɴ] when it precedes [q], and the bilabial [m] before [p]. The vibrant [ɾ] becomes a retroflex sibilant [ʐ] at word onset. The voiced bilabial /b/, dental /d/ and velar /g/, as well as the voiceless bilabial fricatives /ϕ/ and voiceless retroflex /ʐ/, are used as distinct phonemes only in borrowings from Spanish. (MINEDU, 2020).

3.2. Aymara

The Aymara language (ISO 639-3 aym) belongs to the Aru linguistic family, is spoken by the Aymara people and although it is in a vital state (MINEDU, 2018), it is considered an endangered language (Adelaar, 2014). Aymara is spoken in four countries: Argentina, Bolivia, Chile and Peru. In Peru, it is the second most spoken native language after Quechua, according to the 2017 census conducted by the National Institute of Statistics and Informatics (INEI, 2017).

It is an agglutinative language.

Aymara has 3 short phonemic vowels /a/, /i/ and /u/, and 3 long *ä* /aa/, *ï* /ii/ and *ü* /uu/. Also, it features 26 consonant phonemes, most of them aspirated occlusives *ph* [p^h], *th* [t^h] and *kh* [k^h]. In addition, the aspirated postalveolar affricate is signaled by the triplet *chh* [tʃ^h] and an apostrophe is used to signal the occlusive and affricate ejective *p'* [p'], *t'* [t'], *ch'* [ch'] and *k'* [k']. Like Spanish and Quechua it features the phonemes /t͜ʃ/ *ch*, /ɲ/ *ñ*, and /ʎ/ *ll* (MINEDU, 2021).

3.3. Shipibo-Konibo

The Shipibo-Konibo people are one of the most influential communities in the Peruvian Amazon. They call themselves "Jonikon", which means "real people"; they also adopted the exonym "shipibo". Their own language or 'joikon', 'true language' is now known as Shipibo-Konibo. This language belongs to the Panoan linguistic family, which is an important subject of study for many linguistic researchers in Peru (Adelaar, 2014; Zariquiey and others, 2006). Shipibo-Konibo is an agglomerative language, with a high use of common suffixes (130) plus some prefixes (13) for its word-formation process. Furthermore, the basic sentence order is SOV (subject-object-verb) as opposed to Spanish (SVO) (Valenzuela, 2003). This language is spoken by around 22 thousand people in 150 communities and is taught in almost 300 public schools (Sullón Acosta et al., 2013). The majority of the population is bilingual, meaning they speak Shipibo-Konibo and Spanish. Although Shipibo-Konibo is still transmitted to children, there is a growing number of people who speak Spanish as a dominant language and achieve only partial or passive mastery of their native language. Furthermore, the degree of impact of Spanish speech and structure on Shipibo-Konibo is considerable. For these reasons, the language is considered to be in a vulnerable situation.

The phonological repertoire of Shipibo consists of 16 consonants and 4 vowels.

The vowels in Shipibo are characterized by the presence of two heights (high and low), among which it is important to point out the high central vowel, not rounded *ɨ*. In the consonant phonemes we find four labial [p], [b], and [m], nine coronal [t], [s], [ts], [n], [ʃ], [t͜ʃ], [y], [ʂ] and [r], two dorsal [k] and [w] and one global [h] (Martinez, 2009).

4. Corpus Creation

4.1. Methodology

Like Common Voice, we used the crowdsourcing method, which is based on the massive help of volunteers for audio recordings. This methodology allowed us to collect as many audios as possible in a short time and with a small budget. We used two corpus collection applications (Huqariq, Tarpuriq) designed exclusively to record and validate respectively. Unlike the Common Voice platform, the volunteers do not have to read a sentence but listen to it. This last functionality is important for native languages of Peru, due to a large part of the native speaker population are illiterate.

4.2. Text Corpus

This section describes the steps followed to collect the text to be used in the corpus.

The official dictionaries of each language described in this research were used. These dictionaries are publicly available on the Internet. We used the official dictionaries issued by the Peruvian Ministry of Education, because the texts in the dictionaries are correctly written according to the official standard of each language. In Table 1 we can observe the dictionaries used for the creation of our corpus.

In order to organize the data of the collected dictionaries, a table was created manually with the following columns: Language, family, variety, region, author, dictionary name, year, lexical entry, grammatical category, gloss, definition in Spanish, definition in source language, synonym in Spanish, synonyms in source language, notes (clarifications), example in Spanish and example in source language. This table contains all the data from the dictionaries collected. This table was very helpful for the linguists who supported us in the project, since they could make filters to be able to review the data in a simpler and faster way. Finally, the entries that did not have an example in the source language were eliminated, since these examples are used as transcriptions in the corpus.

4.3. Preprocessing and Normalization

After obtaining all the data from the dictionaries of the different languages in a table, we eliminated all the sentences in the ”example in source language” column that had more than 10 words. This was done so that the volunteers would not have problems remembering the sentence to repeat when recording their voices.

Subsequently, 4 native speaker linguists corrected, normalized and standardized the sentences in the ”example in source language” column according to the grammar issued by the Ministry of Education and Ministry of Culture for each language. On the other hand, for the Southern Quechua sentences, a morphological analyzer (Rios, 2015) was used, which automatically standardizes according to the rules of the Ministries of Education and Culture. Table 2 shows the number of sentences we selected from each language.

4.4. Recording of prompts

Linguists who are native speakers of each language recorded their voices reading each of the selected sentences. The recordings were made using the Tarpuriq application for Android, which has an audio recording module very similar to Huqariq application for Android (Camacho and Zevallos, 2020). The recordings were made in a controlled environment, mainly free of noise and interference of any kind. All recordings made by the linguists were stored in a folder called ”prompts” and folders named after their respective languages. All the recordings (prompts) made by the linguists are then entered into the Huqariq application so that they can be listened to by the volunteers to record their voices.

Finally, the prompts were saved as 16-bit, single-channel WAV audio files with a sampling frequency of 16 kHz.

4.5. Recording and validation of audios

For the collection of recordings (audio files) from native speakers (users), Huqariq was used. This application allowed native speakers to record their voices repeating the sentences they hear in the prompts mentioned above. The app assigns 200 sentences per user, this feature of Huqariq was developed in this research in order for users to have a goal and to be able to be rewarded when they achieve it.

The recordings of the volunteers have the same technical information as the prompts. The recordings made by users were validated using 2 methods. The first method used an automated quality validation module that checks the noise, silence and duration of the recordings, this method was incorporated into the Huqariq application. The second method was performed by Tarpuriq, which allowed native linguists of the respective languages to validate the quality of the recordings through a voting system, this method is similar to the one used by Common Voice. Each recording must be voted 3 times, if a recording receives two positive votes, it will be marked as valid, on the contrary, if it receives two negative votes, it will be marked as invalid. Recordings marked as valid will be added to the final training, development and test corpus. These 2 methods allow to have a good quality corpus.

The validated recordings were stored in a folder where they were subsequently divided into three data sets (train, dev, test) according to statistical power analyses. Given the total number of validated recordings in a language, the number of recordings in the test set is equal to the number needed to achieve a 99% confidence level with a margin of error of 1% relative to the number of recordings in the training set. The same is true for the development set.

Table 3 shows the number of hours recorded and validated for each language. As can be seen, Southern Quechua has the highest number of hours collected. This is due to the fact that Southern Quechua has the largest number of native speakers compared to the other languages in this study. In addition, it has a greater participation in revitalization tasks due to the majority of research carried out for this language. Central Quechua, Aymara and Shipibo-Konibo, on the other

Language	Dictionary	Year
Southern Quechua	Yachakuqkunapa Simi Qullqa (Chanka and Collao)	2005
Central Quechua	Chawpi Qichwapa Chimi Qullqan	2017
Aymara	Yatiqirinaka Aru Pirwa	2005
Shipibo-Konibo	Diccionario Shipibo-Español	1993

Table 1: Dictionaries used for the construction of the corpus.

Language	Number of Sentences
Southern Quechua (Chanka and Collao)	8000
Central Quechua	1171
Aymara	1900
Shipibo-Konibo	500

Table 2: Number of sentences for each language used in the construction of the corpus.

Language	Hours		
	Train	Dev	Test
Southern Quechua	144	18	18
Central Quechua	16	2	2
Aymara	12	1	1
Shipibo	4	1	1

Table 4: Statistics of the number of hours divided according to train, dev and test and by each language

hand, unfortunately have little or no participation in revitalization or cultural promotion.

This corpus is a July 2021 version of the corpus, which is the most updated, since due to the pandemic we have not been able to continue working on the validation of the corpus. The corpus currently has a private license, since part of the work was done with funds from private entities. For this reason, those interested can write to us if they wish to make use of it. On the other hand, the corpus statistics are visible on the Siminchikkunarayku[1] page where the following information can be seen: language, phrase, votes, gender, accent. This information is relevant for different types of research and for that reason we consider it useful to place it, as well as in Table 4 the number of hours divided by each language.

Language	volunteers	Hours	
		Total	validated
Southern Quechua	480	340	180
Central Quechua	20	20	20
Aymara	8	15	14
Shipibo	2	7	6

Table 3: Huqariq current data statistics. This data is from the July 2021 version.

5. Automatic Speech Recognition Experiments

The following experiment demonstrates the potential of the Huqariq corpus for multilingual speech research for low-resource languages.

For this experiment we used the corpus described in Table 4. We used the pre-trained model Wav2Vec2 (Baevski et al., 2020) which was trained with 600 hours of Spanish[2]. It is important to mention that we use a pre-trained model of Spanish because the languages in our corpus contain many borrowings from Spanish and this can improve the performance of the model. Moreover, we used the training setup from the public repository of which we obtained the pre-trained model. We trained our models on a GPU with 8 GB of memory for about 24 hours. In addition, we used the Adam optimizer, a learning rate of $4x10^{-5}$ and chose the Wav2letter++ decoder to obtain LM-biased results (Pratap et al., 2019). For the other four languages the modeling units are determined by the BPE algorithm, as in (Zhou et al., 2018). For the experiments, we add an additional projection layer and fit the ASR model with CTC loss as (Yi et al., 2020). During decoding, 5-gram models are used, each of which is trained with the corresponding training transcripts.

Table 5 shows the results of the Wav2Vec2 model for each trained language and the results of previous work. The character error rate (CER) of the resulting model in the test set, defined as the Levenshtein distance (Fiscus et al., 2006) of characters between the true transcription and the decoding result was used to measure the performance of the models for each language. It can be seen from Table 5 that the Wav2Vec2 model does not outperform the previous work for Southern Quechua, this lead to the assumption that the amount of corpus used to train the Wav2Vec2 model is not large enough for the decoder to generalize well. On the other hand, the results for the other languages cannot be compared since in their case for the first time the decoder has been able to generalize well.

[1] www.siminchikkunarayku.pe

[2] https://huggingface.co/facebook/wav2vec2-large-xlsr-53-spanish

Model	Southern Quechua	Central Quechua	Aymara	Shipibo
wav2letter++	31.48	-	-	-
+ (DA)	**22.75**			
Wav2Vec2	28.73	41.15	59.81	72.15
+ CTC (subword)				
+ LM (decode)	23.19	**36.37**	**52.6**	**67.47**

Table 5: Performance results of the ASR models performed for each language using the CER metric.

6. Concluding remarks

We have presented Huqariq: a multilingual speech corpus of Peruvian native languages for the development of speech recognition tools. By using the crowdsourcing methodology and 2 mobile applications we have collected the largest speech corpus of native Peruvian languages. In addition, we have made some modifications to the collection applications so that they are better adapted to the problems of poorly resourced and endangered languages. We are going to release a Creative Commons CC0 licensed version so that the corpus can be in the public domain. On the other hand, we have conducted some experiments on automatic multilingual speech recognition with the Huqariq corpus using the Wav2Vec2 model. This is the first time that speech recognition experiments have been performed for Central Quechua, Aymara and Shipibo-Konibo. Finally, we are working toward the goal that by the end of 2022 Huqariq will be able to work with 20 native languages of Peru and that many more native speakers of these languages will become volunteers.

7. Acknowledgments

We thank all the volunteers who are native speakers of these beautiful languages native to Peru for their time, and especially the volunteer researchers who find new texts to translate and add to the application. Special thanks to Roger Gonzalo, Virginia Mamani, and Abel Anccalle for their work on Huqariq, and all the members of the Siminchikkunarayku team.

This work was supported by the Pontifical Catholic University of Peru in 2020-2021.

8. Bibliographical References

Adelaar, W. F. H. (2014). Endangered languages with millions of speakers: Focus on quechua in peru. *JournaLIPP 3, 2014, 1-12.*

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670.*

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477.*

Camacho, L. and Zevallos, R. (2020). Language technology into high schools for revitalization of endangered languages. In *2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4. IEEE.

Cardenas, R., Zevallos, R., Baquerizo, R., and Camacho, L. (2018). Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP 2*, page 21.

Duan, M., Fasola, C., Rallabandi, S. K., Vega, R. M., Anastasopoulos, A., Levin, L., and Black, A. W. (2019). A resource for computational experiments on mapudungun. *arXiv preprint arXiv:1912.01772.*

Fiscus, J. G., Ajot, J., Radde, N., Laprun, C., et al. (2006). Multiple dimension levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech. In *LREC*, pages 803–808. Citeseer.

Heggarty, P., Valko, M. L., Huarcaya, S. M., Jerez, O., Pilares, G., Paz, E. P., Noli, E., and Usandizaga, H. (2005). Enigmas en el origen de las lenguas andinas: aplicando nuevas técnicas a las incógnitas por resolver. *Revista Andina*, 40:9–57.

INEI. (2017). *Instituto Nacional de Estdistica e Informática.*

Landerman, P. N. (1992). Quechua dialects and their classification. *PhD Thesis.*

Martinez, R. R. (2009). La velarización en shipibo. *Escritura y pensamiento*, 12(24):91–134.

MINEDU. (2018). Documento nacional de lenguas originarias del perú.

MINEDU. (2020). Chawpi qichwata alli qillqanapaq maytu 2= manual de escritura en lengua originaria quechua central.

MINEDU. (2021). Aymara arutha chiqapa qillqañataki panka= manual de escritura aimara.

Pratap, V., Hannun, A., Xu, Q., Cai, J., Kahn, J., Synnaeve, G., Liptchinsky, V., and Collobert, R. (2019). Wav2letter++: A fast open-source speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464. IEEE.

Rios, A. and Mamani, R. C. (2014). Morphological disambiguation and text normalization for southern quechua varieties. *COLING 2014*, page 39.

Rios, A. (2015). *A basic language technology toolkit* *for Quechua*. Ph.D. thesis, University of Zurich.

Rogers, C. and Campbell, L. (2015). Endangered languages. In *Oxford Research Encyclopedia of Linguistics*.

Sullón Acosta, K. N., Huamancayo Curi, E., Mori Clement, M., and Carbajal Solis, V. (2013). Documento nacional de lenguas originarias del perú.

Torero, A. (1964). Los dialectos quechua.

Valenzuela, P. M. (2003). *Transitivity in shipibo-konibo grammar*. University of Oregon.

VoxForge. (2019). Voxforge. In *http://www.voxforge.org/*.

Yi, C., Wang, J., Cheng, N., Zhou, S., and Xu, B. (2020). Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*.

Zariquiey, R. et al. (2006). Reinterpretación fonológica de los préstamos léxicos de base hispana en la lengua. *Boletín de la Academia Peruana de la Lengua*, (41):59–78.

Zhou, S., Xu, S., and Xu, B. (2018). Multilingual end-to-end speech recognition with a single transformer on low-resource languages. *arXiv preprint arXiv:1806.05059*.